Conditional Probabilities, Relative Operating Characteristics, and Relative Operating Levels

SIMON J. MASON AND NICHOLAS E. GRAHAM

International Research Institute for Climate Prediction, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California

(Manuscript received 21 July 1998, in final form 19 April 1999)

ABSTRACT

The relative operating characteristic (ROC) curve is a highly flexible method for representing the quality of dichotomous, categorical, continuous, and probabilistic forecasts. The method is based on ratios that measure the proportions of events and nonevents for which warnings were provided. These ratios provide estimates of the probabilities that an event will be forewarned and that an incorrect warning will be provided for a nonevent. Some guidelines for interpreting the ROC curve are provided. While the ROC curve is of direct interest to the user, the warning is provided in advance of the outcome and so there is additional value in knowing the probability of an event occurring contingent upon a warning being provided or not provided. An alternative method to the ROC curve is proposed that represents forecast quality when expressed in terms of probabilities of events occurring contingent upon the warnings provided. The ratios used provide estimates of the probability of an event occurring given the forecast that is issued. Some problems in constructing the curve in a manner that is directly analogous to that for the ROC curve are highlighted, and so an alternative approach is proposed. In the context of probabilistic forecasts, the ROC curve provides a means of identifying the forecast probability at which forecast value is optimized. In the context of continuous variables, the proposed relative operating levels curve indicates the exceedence threshold for defining an event at which forecast skill is optimized, and can enable the forecast user to estimate the probabilities of events other than that defined by the forecaster.

1. Introduction

Contingency tables are highly flexible methods that can be used to estimate the quality of deterministic and probabilistic forecast systems that express output in continuous, categorical, or binary mode. In their simplest form, contingency tables indicate the quality of a forecast system by considering its ability to anticipate correctly the occurrence or nonoccurrence of predefined events that are expressed in binary terms. For example, precipitation occurrence can be represented on a binary scale by defining an event if precipitation occurred, and a nonevent if there was no precipitation. Data that typically are measured in continuous format can be reduced to a binary statement by, for example, defining whether a season's rainfall occurred in the bottom tercile of climatological seasonal totals. Similarly, the forecast is reduced to a binary statement of whether the defined event is expected to occur. A warning, W, is defined as a forecast of an event, E, occurring. For probabilistic forecasts, warnings are issued when the forecast probability of an event occurring exceeds a predefined threshold.

A two-by-two contingency table can be constructed for a binary system as illustrated in Table 1. From a total number of n observations, the total number of events is given by e, and of nonevents by e'; the total number of warnings is given by w, and of no-warnings by w'. The following outcomes are possible: a hit, if an event occurred and a warning was provided (h is the number of hits); a false alarm, if an event did not occur but a warning was provided (f is the number of false alarms); a miss, if an event occurred but a warning was not provided (m is the number of misses); a correct rejection, if an event did not occur and a warning was not provided (c is the number of correct rejections). The relative (or receiver) operating characteristic (ROC) curve (Swets 1973; Mason 1982; Harvey et al. 1992) is a useful method of representing forecast skill that is based on such a contingency table. In this paper, some guidelines for interpreting the ROC curve are provided, and some of the limitations of the curve are outlined. An alternative way of summarizing the information in the contingency table is examined and compared with the ROC curve.

Corresponding author address: Dr. Simon Mason, IRI, Scripps Institution of Oceanography, University of California, San Diego, Mail Code 0235, La Jolla, CA 92093.
E-mail: simon@lacosta.ucsd.edu

TABLE 1. Two-by-two contingency table for verification of a binary forecast system.

	Forecasts		
Observations	Warning, W	No warning, W′	Total
Event, E	h	m	e
Nonevent, E′	f	c	e'
Total	w	w'	n

2. Data and experimental methods

The ECHAM3[1]-T42 general circulation model was forced using observed sea surface temperatures for the 45-yr period 1950–94. An ensemble of 10 runs was produced, and daily rainfall rates over the 3-month period September–November were averaged for each of the 10 ensemble members over an area representing eastern Africa (10°N–10°S, 30°–50°E). The September–November period constitutes an important rainfall season over eastern Africa, and potentially useful predictability of the rains during this season has been demonstrated elsewhere (Mutai et al. 1998). Rainfall data from an updated version of the Hulme (1992) observational dataset were similarly averaged over the same area. The correlation between the ensemble-mean rainfall and the observations is 0.625, which, over a 45-yr period, provides a strong indication of skill. The simulation provides an indication of potential predictability by estimating the forecast skill that could be achieved using this atmospheric general circulation model, given perfect sea surface temperature forecasts. The procedure of calculating area-averaged rainfall was repeated for the March–May rainfall season, which is considered less predictable than the September–November season (Mutai et al. 1998). The correlation between the ensemble-mean rainfall and the observations is 0.085, over the 45-yr period, effectively indicating no skill.

For each of the two seasons, the 45 yr of observed and simulated area-averaged rainfall were grouped into equiprobable terciles. The three categories are referred to as ''below normal,'' ''near normal,'' and ''above normal.'' The ensemble-mean rainfall was categorized in this way, as well as the simulated rainfall from the individual ensemble members. For each year the percentages of the ensemble members with simulated rainfall in each of the three rainfall categories were calculated.

3. Relative operating characteristic

a. Deterministic systems

The relative operating characteristic is a representation of the skill of a forecast system in which the hit rate and the false-alarm rate are compared (Swets 1973). Both ratios can be calculated simply from the contingency table (Mason 1982):

$$\text{hit rate} = h/(h + m) = h/e = p(\text{W}|\text{E}); \quad (1)$$

$$\text{false-alarm rate} = f/(f + c) = f/e' = p(\text{W}|\text{E}'). \quad (2)$$

The false-alarm rate differs from the false-alarm ratio (Doswell et al. 1990; Schaefer 1990; Harvey et al. 1992), which is traditionally calculated as

$$f/(f + h) = f/w = p(\text{E}'|\text{W}). \quad (3)$$

[There is some inconsistency in the literature, and sometimes no distinction between the false-alarm rate and false-alarm ratio is made. Consequently, both Eqs. (2) and (3) have been referred to as the false-alarm rate (Wilks 1995).] The hit and false-alarm rates, as defined here, fully represent the information in the contingency table.

The hit and false-alarm rates, respectively, indicate the proportion of events for which a warning was provided correctly, and the proportion of nonevents for which a warning was provided incorrectly. The hit rate is sometimes known as the probability of detection, or prefigurance (Olson 1965; Panofsky and Brier 1965; Murphy and Winkler 1987; Doswell et al. 1990; Harvey et al. 1992; Wilks 1995), and provides an estimate of the probability that an event will be forewarned. The false-alarm rate estimates the probability that for a nonevent a warning will be provided incorrectly.

For a system that has no skill, the warnings and events are by definition independent occurrences, and so the probability that a warning was provided is not contingent upon an event occurring or not occurring. In other words, the probability that a warning was provided is unrelated to the outcome and so

$$p(\text{W}|\text{E}) = p(\text{W}|\text{E}') = p(\text{W}). \quad (4)$$

From Eqs. (1), (2), and (4), it is true by definition that when there is no skill the hit and false-alarm rates will both be equal to the prior probability of a warning being provided, $p(\text{W})$ (Murphy and Winkler 1987). This equality occurs when warnings are issued randomly, and when perpetual warnings or no-warnings are provided. When the forecast system has some skill, the hit rate exceeds the false-alarm rate; negative skill is indicated when the false-alarm rate exceeds the hit rate. Because of the equality of the hit and false-alarm rates for all forecast strategies with no skill, the difference between the two ratios could be considered an equitable skill score (Gandin and Murphy 1992). Alternatively, the likelihood ratio, defined as

$$\text{likelihood ratio} = p(\text{W}|\text{E}) \div p(\text{W}|\text{E}'), \quad (5)$$

is close to 1.0 when there is no skill, is larger than 1.0 when there is skill, and approaches 0.0 when there is negative skill. The likelihood ratio indicates how much more (or less) likely it is that a warning was issued before an event than before a nonevent.

[1] ECHAM 3 is version 3.6 of ECHAM, which is from the Max-Planck-Institut für Meteorologie in Hamburg, Germany.

TABLE 2. Area-averaged observations and ensemble-mean simulations of Sep–Nov rainfall over eastern Africa (10°N–10°S, 30°–50°E) for the period 1950–94. The observations and ensemble-mean simulations are expressed in tercile format, with ''B'' representing below-normal, ''N'' near-normal, and ''A'' above-normal rainfall. Also presented are the percentages of the individual ensemble members that simulated rainfall in each of the three categories.

			Probs.						Probs.	
Year	Obs.	Sim.	B	N	A	Year	Obs.	Sim.	B	N	A
1950	B	B	50	40	10	1973	N	N	40	30	30
1951	A	A	10	30	60	1974	B	B	40	60	0
1952	N	N	20	60	20	1975	B	B	60	30	10
1953	N	B	90	10	0	1976	B	N	20	60	20
1954	B	B	100	0	0	1977	A	A	10	30	60
1955	B	B	70	20	10	1978	A	A	0	20	80
1956	N	B	100	0	0	1979	B	B	60	40	0
1957	B	N	20	70	10	1980	N	N	30	30	40
1958	B	N	30	50	20	1981	N	A	20	40	40
1959	N	B	70	30	0	1982	A	A	0	20	80
1960	B	B	80	20	0	1983	B	A	10	20	70
1961	A	A	10	20	70	1984	A	N	40	30	30
1962	N	A	10	30	60	1985	N	B	60	30	10
1963	A	N	20	60	20	1986	N	B	50	30	20
1964	B	B	60	40	0	1987	N	N	30	30	40
1965	A	N	10	40	50	1988	A	A	10	30	60
1966	A	A	10	10	80	1989	A	A	20	40	40
1967	A	A	0	50	50	1990	N	N	10	40	50
1968	A	N	20	60	20	1991	B	N	30	60	10
1969	N	B	40	60	0	1992	B	A	10	20	70
1970	B	N	40	20	40	1993	N	B	50	50	0
1971	N	N	40	20	40	1994	A	A	0	10	90
1972	A	A	0	10	90						

An example is provided in Table 2, where simulations of September–November rainfall over eastern Africa (10°N–10°S, 30°–50°E) using the ECHAM3-T42 model are summarized. The ensemble-mean simulations and the observations have been grouped into three equiprobable categories, and the observed and simulated categories are indicated for each year from 1950 to 1994. A ''B'' represents rainfall in the driest third of the years, ''N'' indicates rainfall in the middle third of the years, and ''A'' represents rainfall in the wettest third. Above-normal rainfall was simulated when above-normal rainfall conditions occurred for 1951, 1961, 1966, 1967, 1972, 1977, 1978, 1982, 1988, 1989, and 1994. The number of hits is therefore 11 (Table 3a), and given that there are 15 events in total, the hit rate is 0.733. The hit rate indicates that 73% of the above-normal rainfall events were simulated correctly over the 45-yr period. In a forecasting environment, a hit rate of 0.733 provides an estimate that a warning could be provided for 73% of future above-normal rainfall events, assuming no change in predictability or forecast performance. In 1962 above-normal rainfall was simulated but did not occur, and so 1962 constitutes a false alarm. There are a total of four false alarms for above-normal rainfall conditions, and given that there are 30 years when rainfall conditions were not above normal, the false-alarm rate is 0.133 (Table 3a). A contingency table for the below-normal category is provided in Table 3b, for which the hit rate is 0.533, and the false-alarm rate is 0.233. The hit rate for above-normal (below normal) rainfall of 0.733 (0.533) is greater than the false-alarm rate of 0.133 (0.233), thus suggesting a high level of skill that could be expected given the strong positive correlation between the ensemble-mean and observed rainfall.

TABLE 3. Contingency tables for the ensemble mean simulation of Sep–Nov rainfall over eastern Africa (10°N–10°S, 30°–50°E) for the period 1950–94. Tables are provided for the simulation of (a) below-normal and (b) above-normal rainfall.

	Forecasts		
Observations	Above normal	Not above normal	Total
a.			
Above normal	11	4	15
Not above normal	4	26	30
Total	15	30	45
b.			
Above normal	8	7	15
Not above normal	7	23	30
Total	15	30	45

The ECHAM3-T42 simulation of March–May rainfall over eastern Africa provides an example of a system with minimal skill. The simulations are summarized in Table 4, and the ensemble-mean contingency tables for above- and below-normal rainfall are presented in Table 5. In the case of above-normal rainfall, the hit rate of 0.267, and false-alarm rate of 0.367 are close to the prior probability of a warning (0.333). The likelihood

TABLE 4. Area-averaged observations and ensemble-mean simulations of Mar–May rainfall over eastern Africa (10°N–10°S, 30°–50°E) for the period 1950–94. The observations and ensemble-mean simulations are expressed in tercile format, with ''B'' representing below-normal, ''N'' near-normal, and ''A'' above-normal rainfall. Also presented are the percentages of the individual ensemble members that simulated rainfall in each of the three categories.

			Probs.						Probs.		
Year	Obs.	Sim.	B	N	A	Year	Obs.	Sim.	B	N	A
1950	N	B	70	20	10	1973	B	A	10	10	80
1951	A	A	10	10	80	1974	A	A	0	40	60
1952	A	N	30	50	10	1975	N	N	60	30	10
1953	B	N	0	70	30	1976	A	B	50	50	0
1954	B	A	10	20	70	1977	N	B	30	60	10
1955	B	B	70	20	10	1978	A	N	40	20	40
1956	N	B	50	40	10	1979	A	A	20	20	60
1957	A	B	40	60	0	1980	N	A	40	20	40
1958	N	N	20	50	30	1981	A	N	20	50	30
1959	B	B	50	40	10	1982	N	A	10	50	40
1960	A	N	30	40	40	1983	B	A	0	30	70
1961	B	A	0	10	90	1984	B	B	100	0	0
1962	N	A	0	10	90	1985	N	N	30	20	50
1963	A	B	70	30	0	1986	N	A	10	30	60
1964	A	A	10	40	50	1987	N	N	30	30	40
1965	B	B	70	20	10	1988	B	B	50	50	0
1966	B	B	60	20	20	1989	A	N	20	50	30
1967	A	N	30	60	10	1990	N	A	10	30	60
1968	A	B	60	30	10	1991	N	N	30	40	30
1969	B	A	20	30	50	1992	B	B	70	30	0
1970	N	A	0	10	90	1993	A	N	20	50	30
1971	N	N	50	30	20	1994	B	B	60	20	20
1972	B	N	30	60	10						

ratio of 0.727 indicates that the probability that a warning was provided is slightly less for when an event occurred compared to when an event did not occur. The likelihood ratio of less than 1.0 is an indication that there is weak negative skill in simulating above-normal March–May rainfall. However, the hit and false-alarm rates for below-normal rainfall are the same as for the September–November rainfall, giving a likelihood ratio of 2.286, which is evidence of positive skill. The low correlation between the ensemble-mean simulated and observed March–May rainfall of 0.085 therefore conceals a suggestion of skill in simulating below-normal rainfall conditions for this season.

TABLE 5. Contingency tables for the ensemble mean simulation of Mar–May rainfall over eastern Africa (10°N–10°S, 30°–50°E) for the period 1950–94. Tables are provided for the simulation of (a) below-normal and (b) above-normal rainfall.

	Forecasts		
Observations	Above normal	Not above normal	Total
a.			
Above normal	4	11	15
Not above normal	11	19	30
Total	15	30	45
b.			
Above normal	8	7	15
Not above normal	7	23	30
Total	15	30	45

b. Probabilistic systems and the ROC curve

For probabilistic forecasts, a warning can be issued when the forecast probability for a predefined event exceeds some threshold (Mason 1979). If, for example, it is decided that a warning is to be issued only when there is at least an 80% confidence that an event will occur, then above-normal conditions for September–November rainfall are indicated in 1966, 1972, 1978, 1982, and 1994 (Table 2). A new contingency table could then be constructed and is indicated in Table 6. Different warning thresholds can be used for the predefined event, and a set of hit and false-alarm rates can then be determined. This set of hit rates is plotted against the corresponding false-alarm rates to generate the ROC curve (Figure 1a). While there are a number of indices for summarizing the performance (Mason 1982), the area under the curve, *A*, is the most commonly used (and simplest to calculate) and has become known as the ROC score.

TABLE 6. Contingency table for the simulation of Sep–Nov rainfall over eastern Africa (10°N–10°S, 30°–50°E) for the period 1950–94. Warnings are issued only when at least 80% of the ensemble members simulate above-average rainfall.

	Forecasts		
Observations	Above normal	Not above normal	Total
Above normal	5	10	15
Not above normal	0	30	30
Total	5	40	45

For deterministic forecasts, an ROC curve can be generated by plotting the hit and false-alarm rate for the forecast system, together with the hit and false-alarm rates obtained for perpetual warnings (for which the hit and false-alarm rates equal 1.0) and no-warnings (for which the hit and false-alarm rates equal 0.0) (Fig. 2).

Because there is skill only when the hit rate exceeds the false-alarm rate, the ROC curve will lie above the 45° line from the origin if the forecast system is skillful and the total area under the curve will be greater than 0.5. A simple transformation of the ROC score can be suggested so its range is from 1.0 (for a perfect forecast system) to −1.0 (for a perfectly bad forecast system), with 0.0 indicating no skill:

$$S = 2 \times (A - 0.5). \quad (6)$$

Alternatively, S can be multiplied by 100 to express the score as a percentage.

c. Interpretation of the ROC curve

In general, for skillful forecast systems, the ROC curve bends toward the top left, where hit rates are larger than false-alarm rates. Where the curve lies close to the diagonal the likelihood ratio is close to 1.0, and the forecast system does not provide any useful information. If the curve lies below the line, negative skill is indicated. To illustrate, the curves for above-normal September–November rainfall are bowed well toward the top left, indicating a high likelihood ratio (Fig. 1a). The relatively poor predictability of above-normal rainfall conditions during March–May is evident from the fact that the curve lies much closer to the diagonal (Fig. 1b) than for September–November rainfall (Fig. 1a). There is, however, some indication of skill in simulating below-normal rainfall during the March–May season.

Near the bottom left of the ROC graphs constructed from the probabilistic information rather than the ensemble mean (Fig. 1), warnings are issued only when a high percentage of the ensemble members simulated above-/below-normal rainfall, and so the number of warnings is small. Toward the top right, the criterion for issuing a warning is relaxed, and so warnings are issued more frequently; the hit rate increases accordingly, but the number of false alarms increases as well. When the individual ensemble members were consistent in simulating above-normal September–November rainfall conditions (with at least 80% of the ensemble members in the upper tercile), then rainfall was below normal on each occasion (no false alarms were issued), and warnings are provided for a third of the events (Fig. 1a). However, if a warning of below-normal rainfall conditions is provided when at least 80% of the ensemble members simulated rainfall amounts in the lower tercile, a few false alarms are issued, and less than 15% of the below-normal rainfall events are successfully indicated. For both above- and below-normal rainfall, too few warnings are issued when the criterion is for at least 80% of the ensemble members to simulate rainfall amounts in the respective tercile. However, a high hit rate is achieved relative to the number of false alarms.

In an application where the cost of a false alarm is prohibitively high, warnings of an event should be issued only when there is high confidence in the event occurring. The ROC curve for the above-normal rainfall indicates that a useful number of wet events potentially could be forewarned successfully, with a minimal threat of a false alarm, if warnings are issued only when there is high confidence. If the cost of a miss, rather than of a false alarm, is prohibitively high, then it would be desirable to increase the number of warnings by relaxing the warning criterion. Issuing more warnings should hopefully ensure that the number of hits is increased at the expense of the number of misses, but with the penalty of issuing more false alarms. The ROC curve is useful in helping to identify an optimum warning criterion, by indicating the trade-off between misses and false alarms. In the example for September–November rainfall (Fig. 1a), all above-normal rainfall events can be forewarned if warnings are issued when at least 20% of the ensemble members simulate rainfall amounts in the upper tercile. However, using such a low level of confidence as a threshold is at the cost of issuing warnings almost half the time when rainfall was not above normal. To ensure that warnings are provided for all below-normal rainfall events, it is necessary to issue a warning when 1 or more of the 10 ensemble members (10%) are in the lower tercile. The consequent high frequency of warnings that would be issued means that over 80% of the time that conditions were normal or above normal, a warning of below-normal rainfall would be issued. Evidently, then, although there is strong evidence of skill at simulating September–November rainfall over eastern Africa, the model is consistently better at simulating wet conditions than dry conditions for the region.

For a probabilistic system, the ROC curve illustrates the varying quality of the forecast system at different levels of confidence in the warning (the forecast probability). It is not necessarily the case that a forecast system demonstrates greatest value at the point at which the likelihood ratio is maximized: instead, each user has a specific cost–loss operating structure, and hence the relative frequencies of hits, false alarms, and misses have to be optimized. The ROC curve can be used in helping to identify this optimum strategy in any specific application (Harvey et al. 1992).

4. Relative operating levels

One limitation with the hit and false-alarm rates is that it is not known in advance whether an event is going to occur, only whether a warning has been given. It is considered of additional value to the user to know what is the probability of an event occurring given that a warning has been provided, $p(\mathrm{E}|\mathrm{W})$ (Murphy and

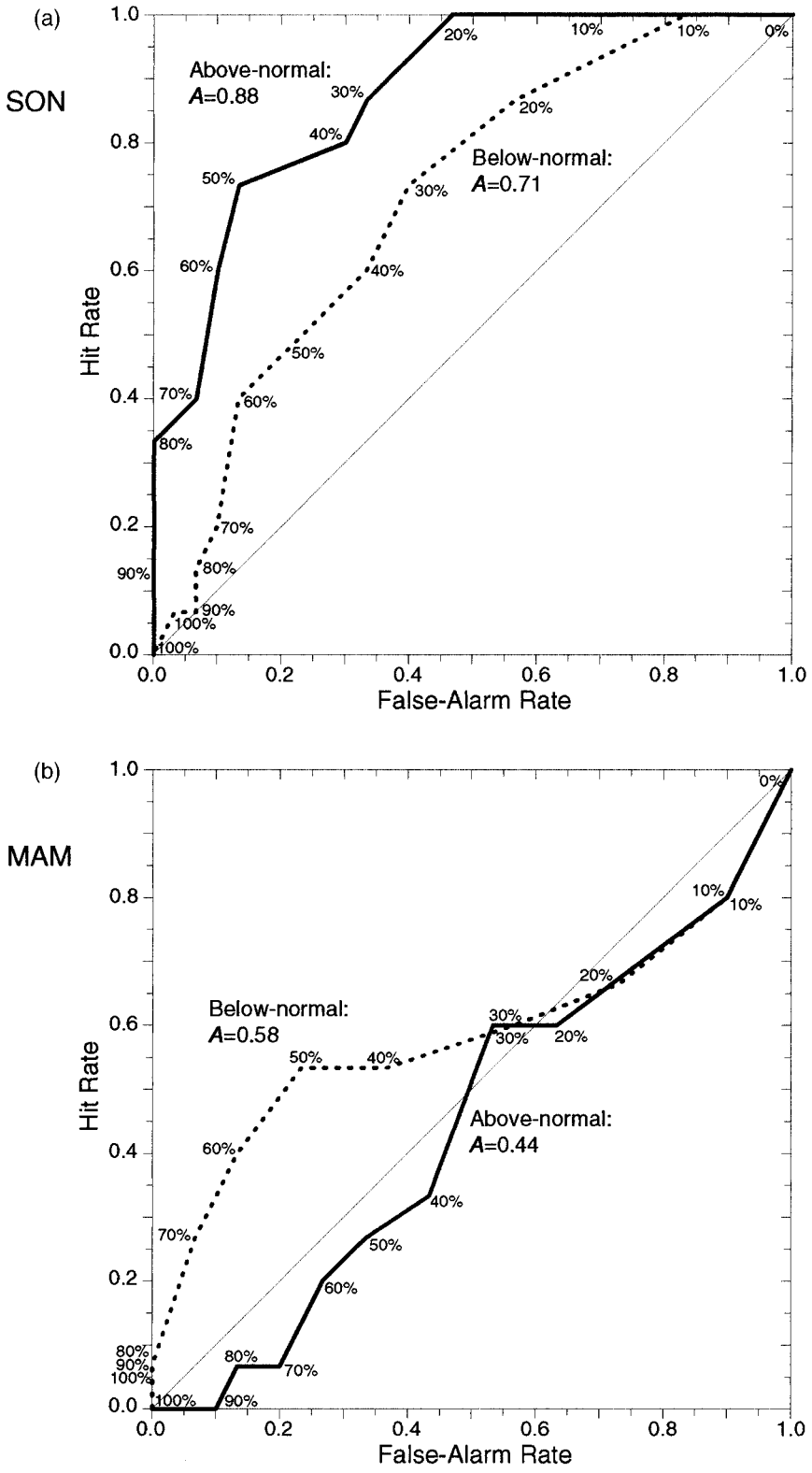

FIG. 1. Hit rates vs false-alarm rates for (a) Sep–Nov and (b) Mar–May area-averaged rainfall for eastern Africa (10°N–10°S, 30°–50°E) from 1950 to 1994. The hit and false-alarm rates were calculated using rainfall simulated by the ECHAM3-T42 general circulation model forced with observed sea surface temperatures and using 10 ensemble members. Results are shown for the simulation of rainfall in the upper (solid line) and lower (dotted line) terciles. Rates are indicated

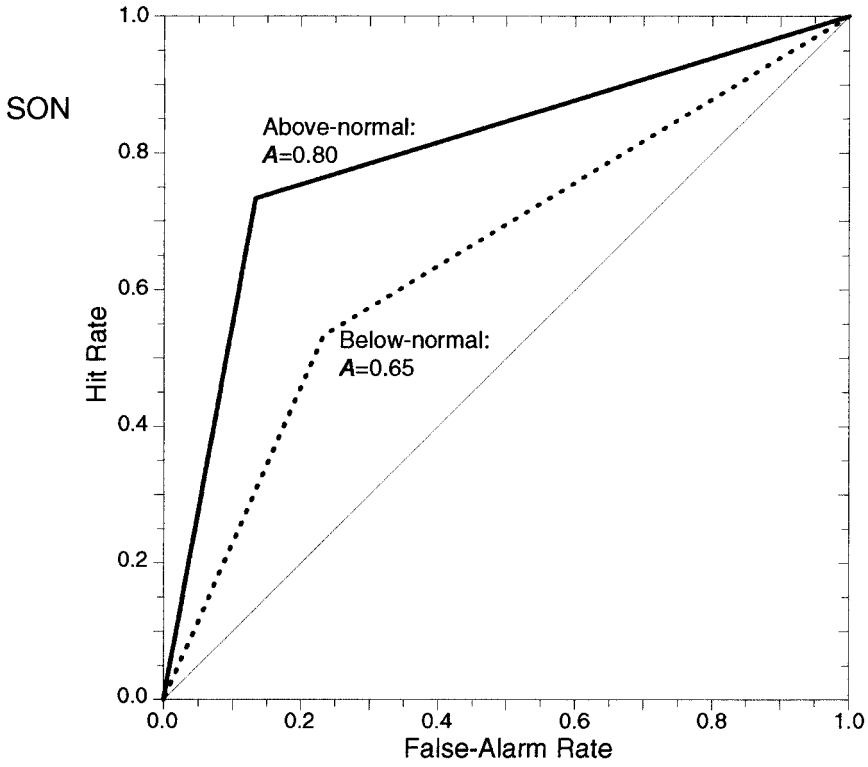

FIG. 2. As for Fig. 1 but for the ensemble-mean simulated rainfall.

Winkler 1987), rather than the probability that a warning was given if an event did occur, $p(\mathrm{W}|\mathrm{E})$. Similarly, it is useful to know what the probability is that an event will occur when a warning has not been provided, $p(\mathrm{E}|\mathrm{W}')$, rather than the probability that a warning was not given if an event did occur, $p(\mathrm{W}'|\mathrm{E})$. [In fact, the false-alarm rate indicates the probability that a warning was given if an event did not occur $p(\mathrm{W}|\mathrm{E}')$, rather than $p(\mathrm{W}'|\mathrm{E})$.] The relationship between $p(\mathrm{W}|\mathrm{E})$ and $p(\mathrm{E}|\mathrm{W})$ is given by Bayes theorem (e.g., Olson 1965):

$$p(\mathrm{E}|\mathrm{W}) = p(\mathrm{W}|\mathrm{E}) \times p(\mathrm{E}) \div p(\mathrm{W}). \quad (7)$$

When training an objective forecast system, the number of warnings frequently is constrained to be equal to the number of events, that is, $p(\mathrm{W}) = p(\mathrm{E})$, in which case it can be shown from Eq. (7) that $p(\mathrm{W}|\mathrm{E}) = p(\mathrm{E}|\mathrm{W})$. In some operational environments, however, the cost of a false alarm relative to that of a miss may be so high that it is preferable to use a strict threshold to define a warning (Harvey et al. 1992), thus constraining warnings to occur infrequently, in which case $p(\mathrm{W}) < p(\mathrm{E})$ and so $p(\mathrm{W}|\mathrm{E}) < p(\mathrm{E}|\mathrm{W})$. Similarly, the cost of a miss may be prohibitive, in which case a low threshold will be used, and $p(\mathrm{W}|\mathrm{E}) > p(\mathrm{E}|\mathrm{W})$. Since $p(\mathrm{E}|\mathrm{W}) \neq p(\mathrm{W}|\mathrm{E})$ in most practical cases, there is some value in a complementary scoring system that is based upon the conditional probability of the outcome given the forecast information (Murphy and Winkler 1987).

a. Correct-alarm and miss ratios

The relative operating levels (ROL) curve is designed to represent the skill of a forecast system from the perspective of the forecasts. It is based upon a calibration-refinement factorization of the contingency table rather than the likelihood-base rate factorization of the ROC (Murphy and Winkler 1987). Whereas the hit rate indicates the proportion of events that were forewarned, the ROL curve makes use of the correct-alarm ratio (Schaefer 1990; Harvey et al. 1992), or postagreement (Olson 1965), which indicates the proportion of times that an event occurred given that a warning had been provided:

$$\text{correct-alarm ratio} = h/(h + f) = h/w = p(\mathrm{E}|\mathrm{W}). \quad (8)$$

The correct-alarm ratio estimates the probability that an

←

using different minimum percentages of ensemble members simulating rainfall in the respective tercile to issue a warning, as indicated by the values along the curves. The areas beneath the curves, *A*, are indicated also.

event will occur given that a warning has been provided. The relationship between the correct-alarm ratio and the hit rate is given by Eq. (7). The ROL curve also makes use of the miss ratio, or detection failure ratio (Doswell et al. 1990). Whereas the false-alarm rate of the ROC score indicates the proportion of nonevents that were forecast as events, the miss ratio indicates the proportion of times an event occurred when no warning had been provided:

$$\text{miss ratio} = m/(m + c) = m/w' = p(\text{E}|\text{W}'). \quad (9)$$

The miss ratio estimates the probability of an event occurring when no warning has been provided. The relationship between the miss ratio [$p(\text{E}|\text{W}')$] and the false-alarm rate [$p(\text{W}|\text{E}')$] is not defined directly by Bayes theorem, but is given by

$$p(\text{E}|\text{W}') = 1 - \frac{[1 - p(\text{W}|\text{E}')]p(\text{E}')}{p(\text{W}')}. \quad (10)$$

For a forecast system that has no skill, the probability of an event is independent of the forecasts (Murphy and Winkler 1987), and so the expected values of the correct-alarm and miss ratios are both equal to the a priori probability of an event occurring:

$$p(\text{E}|\text{W}) = p(\text{E}|\text{W}') = p(\text{E}). \quad (11)$$

This equality occurs when warnings are issued randomly, but not when perpetual warnings or no-warnings are issued. If the probability of a warning approaches 1.0 (0.0) the correct-alarm (miss) ratio tends toward the probability of an event occurring, regardless of the skill of the forecast system. As the probability of a warning tends toward 0.0 (1.0), the correct-alarm (miss) ratio remains determined by the skill of the system, but if perpetual no-warnings (warnings) are provided the correct-alarm (miss) ratio is indeterminable. An important consequence is that, unlike the hit and false-alarm rates, the correct-alarm and miss ratios do not necessarily converge as the probability of a warning tends toward 0.0 or 1.0. Assuming that the probability of a warning is greater than 0.0 and less than 1.0, for a forecast system that has some skill the correct-alarm ratio exceeds the miss ratio (an event is more probable following a warning than a nonwarning); negative skill is indicated when the miss ratio exceeds the correct-alarm ratio.

The maximum possible difference between the correct-alarm and miss ratios (and between the hit and false-alarm rates) occurs when the number of warnings is equal to the number of events (although in practice the maximum observed difference does not necessarily occur at this point). At this point the number of warnings is in a sense ''correct,'' and so it is possible to maximize the number of hits without necessarily issuing false alarms. If the number of warnings is reduced, the miss ratio increases and tends toward the a priori probability of an event as the number of warnings approaches 0.0 (when it can be said that there is no real forecast skill). Similarly, if the number of warnings is increased, the correct-alarm ratio decreases and tends toward the a priori probability of an event as the number of warnings approaches the number of forecasts (when again it can be said that there is no real forecast skill).

As with the hit and false-alarm rates, the correct-alarm and miss ratios can be obtained from forecast summaries, such as Tables 2 and 3. It was shown above that there were 11 hits, when simulating above-normal September–November rainfall, if the ensemble mean is considered (Table 3). Given that above-normal rainfall was simulated for 15 years in total, the correct-alarm ratio is 0.733. (In this case the correct-alarm ratio is the same as the hit rate because the number of warnings and events are identical.) The correct-alarm ratio indicates that 73% of the time above-normal rainfall was simulated, above-normal rainfall occurred. In a forecasting environment, a correct-alarm ratio of 0.733 provides an estimate that there is a probability of 73% that rainfall will be above normal, if a warning of above-normal rainfall is issued. In 1963 above-normal rainfall occurred but was not simulated (Table 2), and so 1963 constitutes a miss. There are a total of four misses for above-normal rainfall conditions, and given that there are 15 years when rainfall conditions were above normal, the miss ratio is 0.267. The miss ratio indicates that 27% of the time above-normal rainfall was not simulated, above-normal rainfall occurred. In a forecasting environment, a miss ratio of 0.267 provides an estimate that there is a probability of 27% that rainfall will be above normal, if a warning of above-normal rainfall is not provided.

Although the correct-alarm ratio provides an estimate of the probability that an event will occur, it is not the case that the correct-alarm ratio is equal to the forecast probability, even in a well-calibrated forecast system. The forecast probability should ideally be equal to the observed relative frequency of an event (Murphy and Winkler 1977), but the correct-alarm ratio provides an estimate of the probability of an event given a forecast probability equal to *or greater than* the current probability. In a well-calibrated system, therefore, the correct-alarm ratio will be greater than the forecast probability. However, before examining the correct-alarm and miss ratios for probabilistic forecasts in greater detail, it is worth considering some features of the graphical presentation of the ratios.

b. The ROL curve

The ROC score is usually calculated by plotting the hit rates against the false-alarm rates for different warning criteria, and then calculating the area under the curve, as discussed above. If the correct-alarm ratio is plotted against the miss ratio in the same way as for the ROC curves, the points of zero skill at different forecast probabilities would be represented on the graph by a single point, since the ratios become a function of the prior probability of the event [Eq. (11)] rather than of

the forecast probability [Eq. (4)]. Other peculiarities in plotting the correct-alarm ratio against the miss ratio for different forecast probabilities arise. First, because there is a direct inverse relationship between the two ratios, it is not possible for both the correct-alarm and the miss ratios to be greater or less than the a priori probability of an event, and so not all points on the graph are feasible. A second feature of the plot is that because of the lack of convergence of the correct-alarm (CAR) and miss ratios (MR) when perpetual warnings or no-warnings are issued, the end points of the curve would always lie on MR = e/n and CAR = e/n, rather than at the origin and at the top right.

Because of these differences in behavior of the ratios used in the ROL curve from those used in the ROC curve, it is inappropriate to calculate the ROL score in a way that is directly similar to that for the ROC score. A possible alternative would be to plot the correct-alarm and miss ratios against the warning probability. Such a plot would be similar to the attributes and reliability diagrams (Hsu and Murphy 1986; Wilks 1995), but includes the miss ratios for additional information. The ROL score could then be calculated as the area between the two curves, counting area as negative where the correct-alarm ratio is less than the miss ratio. An area of greater than zero would be a necessary condition for skill. However, there remains the problem that the correct-alarm and miss ratios do not converge as the forecast probability tends toward 0.0 or 1.0, and so an alternative approach is considered preferable.

A fundamentally different solution would be to fix the definition of the warnings and to calculate the correct-alarm and miss ratios for different events. If the ROL curve is constructed as suggested, it can be demonstrated that the correct-alarm ratio now behaves in a manner identical to that of the hit rate when the forecast probability is varied. Similarly, the miss ratio behaves in a manner identical to that of the false-alarm rate (Fig. 3). Consequently, if the correct-alarm ratios are plotted against the miss ratios for different prior probabilities, the curve converges on the origin and the top right; all points on the graph are valid; the diagonal represents the line of zero skill; skill is indicated when the curve is bowed toward the top left, where the correct-alarm ratio exceeds the miss ratio. Because the correct-alarm and miss ratios display the required properties when plotted with fixed forecast probability and varying a priori probability of an event, the area beneath the curve could be calculated to give an ROL score. Equation (6) could be used to transform the area into a skill score in an identical manner to that for the ROC score.

c. Interpretation of the ROL curve

In constructing the ROC curve the prior probability of an event is kept fixed and the threshold-forecast probability is altered. When forecasting continuous data (such as temperatures or rainfall amounts, as opposed to rainfall occurrence), in many cases the event is defined arbitrarily. In seasonal climate forecasting, for example, rainfall forecasts frequently are presented in the form of probabilities of seasonal rainfall totals being within each climatological tercile, as in the examples shown in this paper, but quintiles or alternative thresholds could be used equally justifiably (Ward and Folland 1991). In an imperfect model or forecast system, in which forecasts of the values of continuous variables are being provided, there is no guarantee that the forecasts provided do not give more reliable indications of the probability of an event different from that specified. For example, the number of ensemble members in the driest tercile may provide better estimates of the probability of precipitation being in the driest quintile rather than the driest tercile. Alternatively, in a forecast system that consistently underforecasts precipitation occurrence, warnings of precipitation occurrence may give reliable estimates of precipitation occurrence of a few millimeters. Two questions then arise: given a predefined event in the observations, what is the optimal definition of this event in the model data; and given a predefined event in the model data, what is the optimal definition of this event in the observations? The latter question is addressed in the ROL curve by varying the definition of an event in the observations so that it is not kept coincident with the definition of an event in the model.

When constructing the ROL curve, the definition of a warning is kept fixed. The minimum forecast probability and the definition of an event in the model data therefore are predefined. If events are measured using continuous data, the definition of an event in the observations can be varied by adjusting the threshold or, equivalently, the prior probability. For example, precipitation occurrence could be defined using different minimum precipitation amounts, or cold spells could be defined using different temperature thresholds. In the example for September–November (Fig. 3a), warnings have been defined when 70% of the ensemble members simulate rainfall amounts in the upper/lower tercile. Near the bottom left of the graph an event is defined when the observation is above (below) the 90th (10th) percentile, implying that the number of events is small. Toward the top right, the criterion for defining an event is relaxed, and so more events are defined.

The ROL curve is able to provide an indication of the estimated probabilities of different outcomes given the forecast criteria. For example, given a warning of a wet event, as defined above, there is an estimated 60% chance of an event beyond the 80th percentile (in the wettest 20% of years) occurring, and an estimated 100% chance of an event beyond the 70th percentile (in the wettest 30% of years). From a warning of dry conditions, there is an estimated 50% chance of an event drier than the 30th percentile, and an estimated 100% chance of an event drier than the 60th percentile. Again there is shown to be less information in the forecasts of dry

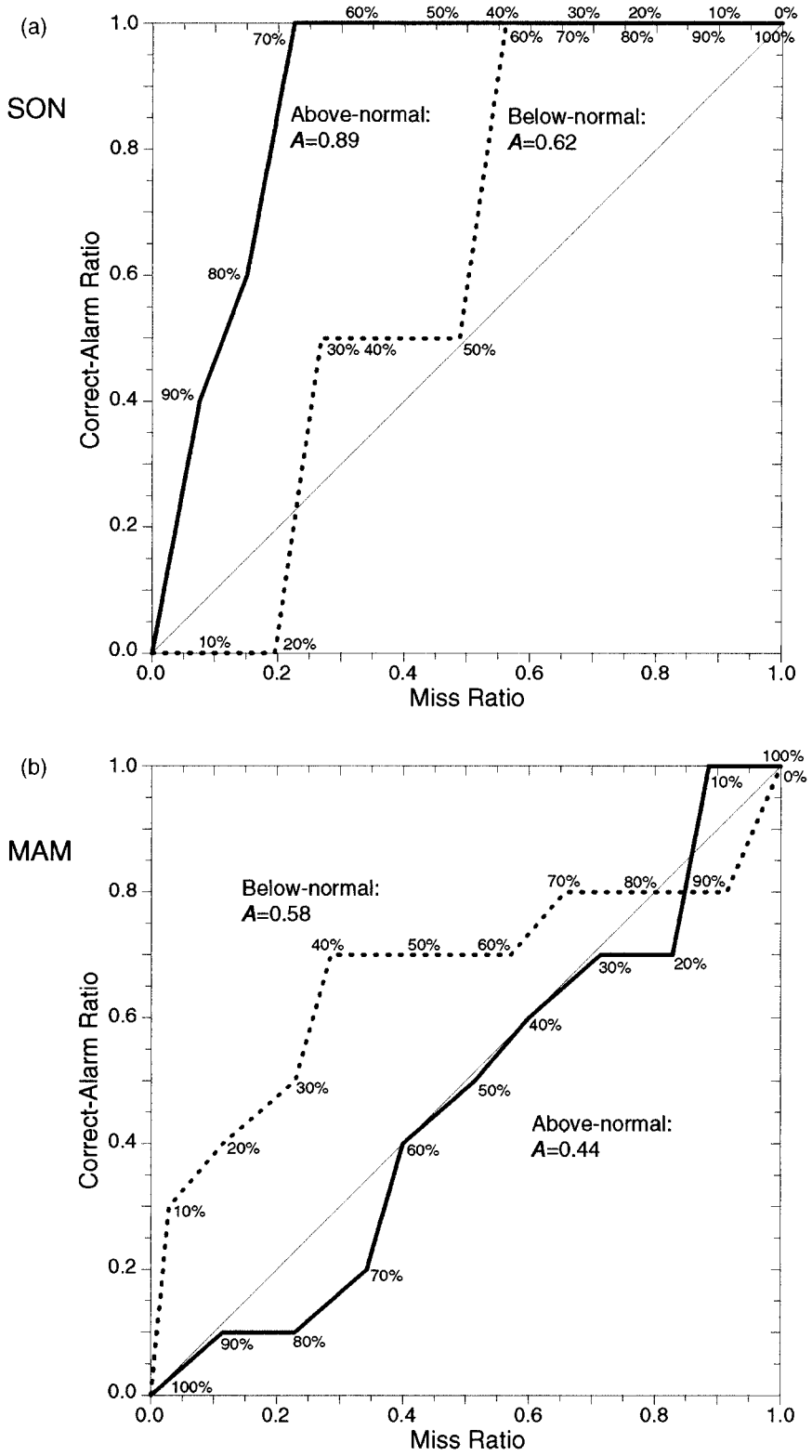

FIG. 3. Correct-alarm ratios against miss ratios for (a) Sep–Nov and (b) Mar–May area-averaged rainfall for eastern Africa (10°N–10°S, 30°–50°E) from 1950 to 1994. The correct-alarm and miss ratios were calculated using rainfall simulated by the ECHAM general circulation model forced with observed sea surface temperatures and using 10 ensemble members. Results are shown for the simulation of rainfall in the upper (solid line) and lower (dotted line) tails. Warnings were

conditions than of wet. Equally it is possible to estimate from the ROL curves the probability of an event given no warning. Given a warning of wet conditions, there was an estimated 60% chance of an event beyond the 80th percentile (in the wettest 20% of years) occurring, while there is an estimated 15% chance of an event given no warning (a miss).

Given the information in the ROL curve, the forecaster is able to obtain some information about the relative predictability of events of differing magnitudes. Similarly, users, for whom access to additional forecast information frequently is restricted, would be able to identify the probabilities of events defined using different thresholds that may be of more direct relevance to their interests than the definition provided by the forecaster. Figure 3b illustrates ROL curves for March–May eastern Africa rainfall, when a warning is issued given that at least 50% of the ensemble members simulate rainfall totals in the above- and below-normal terciles. A slightly weaker criterion is used to define a warning than for the September–November rainfall (Fig. 3a) since the ensemble variance is larger for March–May than for September–November.

When a warning of below-average rainfall was provided, 70% of the time observed rainfall was drier than the 40th percentile, while when no warning was issued, the observed rainfall was drier than the 40th percentile only about 30% of the time. The warnings are designed to indicate a high probability (50%) of rainfall being within the driest third of years, but apparently give a good indication of enhanced probabilities of rainfall totals being in the driest 40% of years. It seems probable that on a number of occasions when warnings of anomalously dry conditions were issued, the observed rainfall was slightly wetter than the climatological third. Perhaps of greater interest, however, is that the warnings give good indications of rainfall totals being in the driest 10% of years. About 30% of the time a warning was issued, a one-in-ten-year drought (defined simply as anomalously low rainfall) occurred, while when no warning was issued droughts of such severity occurred less than 5% of the time. The ROL curve suggests that there may be a useful degree of predictability of extremely dry conditions, as well as of the one-in-three-year droughts of which the warnings were designed to provide an indication. The forecast user is able to obtain estimates of the probabilities of droughts of differing severity, while the forecaster is encouraged to investigate the predictability of extremely dry conditions further.

5. Discussion and summary

The relative operating characteristic, ROC (Mason 1982; Stanski et al. 1989; Harvey et al. 1992), is being considered by the World Meteorological Organization as a recommended method of indicating the skill of probabilistic weather and climate forecasts. The ROC is a highly flexible system that can be used to assess the skill level of dichotomous, categorical, continuous, and probabilistic forecasts. It is based on a 2×2 contingency table and compares the proportion of events that were forewarned (the hit rate) with the proportion of nonevents that occurred after a warning (the false-alarm rate). Given an ensemble of forecasts, it is useful to construct an ROC curve showing different combinations of hit and false-alarm rates given different forecast probabilities. The ROC curve is useful for identifying an optimal strategy for issuing warnings, by indicating the trade-off between false alarms and misses.

In an operational environment, the warning is provided in advance of the outcome, and so there is additional value in knowing the probability of an event occurring, contingent upon the forecast probability. An alternative summary of the 2×2 contingency table is proposed indicating the probabilities associated with different events given a warning (correct-alarm ratio), and the probabilities given no warning (miss ratio). The correct-alarm and miss ratios are useful in estimating probabilities of events from an ensemble of forecasts, which may not provide reliable probabilities because of model biases and errors.

If the correct-alarm and miss ratios are plotted in the same manner as for the ROC curve, peculiarities in the joint behavior of the two ratios impose unwanted constraints on the graph. The recommended alternative is to calculate different values of the two ratios for a fixed definition of a warning, but for varying event definitions. The forecast probability and definition of an event in the model data are kept fixed, but the definition of an event in the observations is varied. The incompatibility of the definition of the event between the model data and the observations can provide valuable information to both the forecaster and the forecast user.

Careful interpretation of the ROC and ROL curves provides a wealth of information about the forecast system. The ROC curve illustrates the varying quality of the forecast system at different levels of confidence in the warning (the forecast probability) and can be used to optimize forecast value given the specifics of an individual user's cost–loss table. The ROL curve can be used by the forecaster to help compare levels of pre-

←

issued when at least (a) 70%, (b) 50% of the ensemble members were in the upper (lower) tercile, and events were defined when the observations were above (below) the percentile indicated by the values along the curves. The areas beneath the curves, *A,* are indicated also.

dictability of events of differing magnitude, and by the forecast user to estimate probabilities of events other than that defined by the forecaster. Because of model biases and systematic errors, forecasts of continuous variables may provide more reliable indications of the probability of events different from that specified. For example, forecasts of precipitation occurrence, in which the forecast is for at least a trace of precipitation to be recorded, may provide more useful information about the probability of at least 5 mm of precipitation, or some other amount. Similarly, given a history of wind speed advisories, a forecast user may be able to estimate the probabilities of different wind speeds given a current forecast. In the context of seasonal climate forecasting, the number of ensemble members in the driest tercile may provide better estimates of the probability of precipitation being in the driest quintile rather than the driest tercile.

The area under the ROC and ROL curves is a simple index for summarizing the skill of a forecast system, but is sensitive to the number of points that are plotted. It has been recommended that for the ROC curve normal probability axes be used instead of linear axes to minimize the effects of the number of points (Mason 1982). This transformation is based on the assumption that the variable used in the decision criterion has a normal distribution prior to an event and prior to a nonevent. Departures from this assumption are not usually large (Mason 1982) and significantly affect the interpretation of scores based on the nonlinear axes only when departures from normality are extreme or sample sizes small (Hanley 1988). Nonparametric methods have been developed for comparing ROC curves, and for testing the significance of individual curves (Centor 1991), but require a minimum number of points. Further discussion of the significance and comparison of ROC or ROL scores is beyond the scope of this paper. However, the effects of sample size and the number of points on the graph should be borne in mind when comparing areas under ROC or ROL curves on linear axes for different forecast systems.

Ideally, the ROC curve should be constructed using probabilistic forecasts so that it is possible to vary the definition of when a warning is issued, based on varying levels of forecast confidence. Nevertheless, it is possible to construct an ROC curve given deterministic forecasts using perpetual warnings and no-warnings as the end points (Fig. 2). Unlike the ROC curve, the number of points on an ROL curve is not constrained when deterministic forecasts are issued. Given deterministic forecasts, there is no freedom to adjust the level of confidence at which a warning is issued, but since the definition of a warning is kept fixed on an ROL curve, the restriction is irrelevant. In fact, the correct-alarm and miss rates of the ROL curve are one means of converting a deterministic forecast into a probabilistic format, and of estimating the reliability of probabilistic forecasts. Instead the ROL curve is constrained when forecasts of dichotomous variables are considered. Ideally an ROL curve requires forecasts of continuous variables, so that the definition of an event can be varied, but when forecasting dichotomous variables it is possible to define the end points of an ROL curve as perpetual no-events (bottom left) and perpetual events (top right).

To illustrate the utility of the ROC and ROL curves, simulations of rainfall over eastern Africa, using the ECHAM-T42 general circulation model forced with observed sea surface temperatures, have been examined. The ROC curves for eastern African rainfall have shown that the ECHAM-T42 model is able to simulate more successfully above-normal September–November rainfall conditions than below normal. Such information is valuable in that it enables the forecaster to provide higher levels of confidence in forecasts of above-normal rainfall for this time of year. Similarly, the ROC curve for the March–May rains, which are generally considered fairly unpredictable, suggests that the model may be able to simulate below-normal rainfall conditions successfully and that there may be predictability when there is a high level of consistency among the ensemble members.

The ROL curves for eastern African rainfall have confirmed that the ECHAM-T42 model is able to simulate successfully below-average rainfall conditions for the March–May season, and suggest that there may be a high level of skill in simulating exceptionally dry conditions. The possibility that there may be some predictability of the March–May rainfall was not indicated by the correlation coefficient, but is suggested by both the ROC and the ROL curves.

Acknowledgments. This paper was funded in part by a grant/cooperative agreement from the National Oceanic and Atmospheric Administration (NOAA). The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its subagencies. The comments of two anonymous referees were of considerable value.

REFERENCES

Centor, R. M., 1991: Signal detectability: The use of ROC curves and their analyses. *Med. Decis. Making,* **11,** 102–106.

Doswell, C. A., III, R. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting,* **5,** 576–585.

Gandin, L. S., and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.,* **120,** 361–370.

Hanley, J. A., 1998: The robustness of the ''binormal'' assumptions used in fitting ROC curves. *Med. Decis. Making,* **8,** 197–203.

Harvey, L. O., K. R. Hammond, C. M. Lusk, and E. F. Mross, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.,* **120,** 863–883.

Hsu, W., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting,* **2,** 285–293.

Hulme, M. H., 1992: A 1951–80 global land precipitation climatology for the evaluation of general circulation models. *Climate Dyn.,* **7,** 57–72.

Mason, I., 1979: On reducing probability forecasts to yes/no forecasts. *Mon. Wea. Rev.*, **107,** 207–211.

——, 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30,** 291–303.

Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115,** 1330–1338.

Mutai, C. C., M. N. Ward, and A. W. Colman, 1998: Towards the prediction of the East Africa short rains based on sea-surface temperature-atmosphere coupling. *Int. J. Climatol.*, **18,** 975–997.

Olson, R. H., 1965: On the use of Bayes' theorem in estimating false alarm rates. *Mon. Wea. Rev.*, **93,** 557–558.

Panofsky, H. A., and G. W. Brier, 1968: *Some Applications of Statistics to Meteorology*. The Pennsylvania State University, 224 pp.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5,** 570–575.

Stanski, H. R., L. J. Wilson, and R. Burrows, 1989: Survey of common verification methods in meteorology. WWW Tech. Rep. 8, WMO/TD 358, 114 pp.

Swets, J. A., 1973: The relative operating characteristic in psychology. *Science*, **182,** 990–1000.

Ward, M. N., and C. K. Folland, 1991: Prediction of seasonal rainfall in the north Nordeste of Brazil using eigenvectors of sea-surface temperatures. *Int. J. Climatol.*, **11,** 711–743.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.